

# **PHYSTAT05**

Two approaches to statistical inference

D.R. Cox

*Nuffield College, Oxford*

david.cox@nuf.ox.ac.uk

## The issue

Data  $y$ . Probability model representing the generation of  $y$  as observed value of random vector  $Y$ . Unknown parameter  $\theta = (\psi, \lambda)$ . Use data to

- draw conclusions about  $\psi$
- criticize or improve model

In this probability is an (idealized) frequency. Parameter  $\psi$  should capture important properties of system free from the accidents of the specific data

How should we express our uncertainties about  $\psi$ ?

## Two broad routes to an answer

- Frequentist
  - keep to a view of probability as a (hypothetical) frequency
  - use this to calibrate procedures which, while not expressing uncertainties directly as probabilities, have appealing properties under hypothetical repetition
- Inverse probability (Bayesian)
  - if necessary extend or change the notion of probability so that we can make probability statements about  $\psi$

## Preliminaries 1

Possible data,  $D_1, D_2, \dots$

Possible explanations  $E_1, E_2, \dots$

$$\begin{aligned}P(E_k | D_j) &= P(E_k \cap D_j) / P(D_j) \\ &= P(D_j | E_k) P(E_k) / P(D_j) \\ &\propto P(D_j | E_k) P(E_k).\end{aligned}$$

Posterior probability  $\propto$  Likelihood  $\times$  Prior probability.

## Preliminaries 2

Essentially the same for parameterized statistical models.

Model specifies  $f_Y(y; \theta)$ , for a given  $y$  called the *likelihood* as a function of  $\theta$ . Prior density specifies marginal density of  $\Theta$  as  $f_\Theta(\theta)$ , as what is known without data  $y$ . Then

$$f_{\Theta|Y}(\theta | y) \propto f_{Y|\Theta}(y | \theta) f_\Theta(\theta).$$

To obtain posterior of  $\psi$  integrate out  $\lambda$ .

## Brief history

- Laplace (flat priors)
- Gauss
- Boole et al
- K.Pearson
- R.A. Fisher (1922)
- J. Neyman and E.S.Pearson (1928-)

Two rather different approaches to frequentist discussions

## **Brief History, ctd**

### Bayesian (impersonal)

- J.M. Keynes
- H. Jeffreys

### Bayesian (personalistic)

- F.P. Ramsey
- B.de Finetti
- I.J. Good
- L.J. Savage

## Outline of frequentist approach

- in most situations frequency-based probabilities cannot be attached to  $\psi$
- use instead measures of security,  $p$ -values and confidence sets in particular, calibrated by what happens when *hypothetically* they are applied repeatedly
- confidence sets via all values of  $\psi$  consistent with data up to a specified level
- may be null or whole space

## **Critique of frequentist approach**

- how do we ensure that the long-run procedure used in calibration is relevant to the analysis at hand?
- technically exact solutions available only for relatively limited class of problems
  - in other cases some kind of approximation, usually based on asymptotic analysis is needed

## **Critique of frequentist approach ctd**

But

- gives a systematic base tied to real world for a wide range of methods not depending on additional specification
- gives an approach for assessing methods that are not necessarily fully efficient in some technical sense

## A simple example

Suppose that  $Y$  has Poisson distribution with mean  $(\gamma + \lambda)t_S$  and that  $Y_B$  has a Poisson distribution of mean  $\lambda t_B$ .

- if interest lies in  $\psi^* = \gamma/\lambda$ , exact efficient solution is provided from binomial distribution of  $Y$  given  $Y + Y_B = t$  which has parameter  $t_S\psi^*/(t_S\psi^* + t_B)$
- but if interest lies in  $\psi = \gamma$  there is no comparable procedure

## An approximate solution

For given  $y$  let  $p(y, \psi_0; \lambda)$  be the  $p$ -value for testing  $\psi = \psi_0$ , assuming  $\lambda$  known. Let  $\tilde{\lambda}$  be unbiased for  $\lambda$  with small variance  $v(\lambda)$ , all conditional on  $y$ . Then a close approximation to the significance level is  $p(y, \psi_0; \lambda^*)$ , where

$$\lambda^* = \tilde{\lambda} - \frac{v(\tilde{\lambda})\partial^2 p/\partial\lambda^2}{2\partial p/\partial\lambda},$$

where the final term may be evaluated at  $\lambda = \tilde{\lambda}$ .

In particular, if  $y = 0$  the  $p$ -value for testing  $\psi = \psi_0$ , leading to an upper confidence limit for  $\psi$ , is

$$\exp\left\{-\left(\psi_0 + y_B/t_B\right)t_S + y_B t_S^2/(2t_B^2)\right\}.$$

## Flat priors and Bayesian methods

It seems to be agreed from all theoretical standpoints that flat or ignorance priors are dangerous, although they are widely used in applications.

- if  $\theta$  has flat (uniform) distribution  $e^\theta$  has an exponential distribution
- can saying that values outside a (possible long) interval are vastly more likely than those inside really be expressing ignorance
- in one dimension Jeffreys prior appealing and also has good frequentist properties
- flat priors in several dimensions may produce clearly unacceptable answers

## Reference priors

Find prior weighting function that maximizes expected discrepancy between prior knowledge and perfect knowledge obtained by replication of system. In finite case leads to maximum entropy prior.

- yields Jeffreys prior in one-dimensional cases
- complicated in multidimensional problems
  - parameters must be ordered in importance
  - priors for a particular parameter change within a given model
  - prior model and design-specific
  - operational interpretation of posterior probability of, say, 0.3 unclear

## A simple example

Suppose we have  $m$  observations from a normal distribution of mean  $\mu$  and  $n$  observations from a normal of mean  $\nu$ , all observations having unit variance. Then if we are interested in  $\psi = \mu\nu$  the reference prior is proportional to

$$(m\mu^2 + n\nu^2)^{-1/2}.$$

This depends on  $m/n$  and does not yield the prior for  $\mu$  on ignoring  $\nu$ .

## **Personalistic probability**

A radically different approach aims to introduce into the analysis uncertain information of more general kind than is represented by statistical data in the narrower sense. In theoretical discussion it is usually set out as a theory of personal decision making. And that is how it is best regarded.

## Personalistic probability, ctd

Suppose for simplicity given a procedure for generating a random event with any specified probability  $p$ . Then  $P(\mathcal{E} \mid \mathcal{F})$ , Your probability of  $\mathcal{E}$  given  $\mathcal{F}$  is found as that  $p$  such that You are indifferent as between

- a valuable prize if  $\mathcal{E}$  is true and zero otherwise
- same valuable prize if event with probability  $p$  occurs and zero otherwise

A certain kind of consistency (coherence) requires laws of probability to hold.

## Personalistic probability, ctd

- not a theory of psychological probability
- no assumption that You and You\* have the same  $p$  given  $\mathcal{F}$

## How can priors be found and what might posteriors mean?

- prior might be a frequency
- prior might be a summary of previous data but then there is an illogicality
- prior might be a reference prior in which case posterior could be interpreted as an approximate confidence interval or possibly directly, although how not clear
  - naive reference priors potentially dangerous
- prior could be an injection of evidence external to the data
  - what is the evidence-base?

## Summary

- formal inferential aspects often very small part of statistical analysis
- carefully used frequentist approach yields broadly applicable if sometimes clumsy answers
- in simple problems appropriate flat priors yield essentially same answer
- to inject further information quantitatively, informative prior might be useful
  - essential to look at evidence-base
  - and probably do sensitivity analysis

- except for personal decision-making general personalistic theory  
inappropriate
- flat priors in general suspect